ChatGPT を用いた臨床検査技師国家試験の正答率の評価

～GPT3.5 から GPT-4o までの比較分析～

◎加藤 洋平[1)]、土井 洋輝[2)]、石田 秀和[1)]、坪井 良樹[2)]、佐々木 健太[3)]、大島 康平[1)]、菊地 良介[1)]
岐阜大学医学部附属病院 検査部[1)]、藤田医科大学 医療科学部[2)]、岐阜大学医学部附属病院 病理部[3)]

【はじめに】近年、大規模言語モデル（large language models; LLM）が様々な分野で世界的に注目を集めている。本研究では OpenAI 社が開発した LLM による対話型 AI サービスである ChatGPT について、過去 3 年間の臨床検査技師国家試験における正答率の評価を行った。

【方法】厚生労働省ホームページに公開されている 2020 年（第 67 回）から 2022 年（第 69 回）の 3 年間の臨床検査技師国家試験問題のうち、画像問題（計 100 問）を除く 500 問を解析対象とした。LLM は GPT-3.5、GPT-4、GPT-4o の 3 モデルを用いた。問題文及び選択肢は改変せず入力し、各モデルを用いて解答を生成した。また、GPT-4o では画像問題に対しても解答を生成し評価した。各モデルの正答率の比較は McNemar’s test を用いた。また、p 値は bonferroni 法にて補正した。画像問題とそれ以外の問題の正答率比較には Chi-squared test を用いた。

【結果】臨床検査技師国家試験における文章問題の正答率は GPT-3.5: 51.4%（257/500 問）、GPT-4: 79.8%（399/500 問）、GPT-4o: 89.2%（446/500 問）であり、3 つのモデル間でいずれも有意な差を認めた（$p<0.001$）。特に、GPT-3.5 は合格基準（60.0%）に満たないのに対して、GPT-4 および GPT-4o は合格基準に達した。また、GPT-4o による画像問題の正答率は 60.0%（60/100 問）であり、文章問題の正答率（89.2%）と比較して有意に低い正答率であった（$p<0.001$）。しかし、文章問題と画像問題を合わせた正答率は 84.3%（506/600 問）と合格基準を上回る正答率だった。

【考察・結語】GPT-3.5 に対して GPT-4.0、GPT-4o において正答率が向上した要因に事前学習量とパラメータ数の増加に伴う精度向上が考えられる。一般的に LLM のような深層学習モデルはパラメータ数が多いほど予測精度が向上すると言われている。GPT-4、GPT-4o のパラメータ数は現在非公開であるが, これまでの GPT におけるパラメータ数の推移からもパラメータ数の大幅な増加が予想される。本結果から ChatGPT に代表される LLM における課題抽出は臨床検査領域での多様な応用進展に寄与することが考えられる。

連絡先 058-230-7251